Constructing local Cell Specific Networks from Single Cell Data

Xuran Wang[1], David Choi[2] and Kathryn Roeder[1,3,*]

March 2021

[1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA,

[2] Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[3]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

[*] To whom correspondence should be addressed. Email: roeder@andrew.cmu.edu

Abstract

Gene co-expression networks yield critical insights into biological processes, and single-cell RNA sequencing provides an opportunity to target inquiries at the cellular level. However, due to the sparsity and heterogeneity of transcript counts, it is challenging to construct accurate gene networks. We develop an approach, locCSN, that estimates cell-specific networks (CSN) for each cell using a method inspired by Dai et al. (7). LocCSN is based on a non-parametric investigation of the joint distribution of gene expression, hence it can readily detect nonlinear correlations and it is more robust to distributional challenges. Although individual CSNs are estimated with considerable noise, average CSNs provide stable estimates of network structure, which provide better estimates of gene block structure than traditional measures. The individual networks preserve the cellular heterogeneity and repeated estimates of network structure facilitate testing for differences in network structure between cell groups. Additionally, we propose new downstream analysis methods using CSNs to utilize more fully the information contained within them. Finally, to further our understanding of autism spectrum disorder we examined the evolution of gene networks in fetal brain cells and compared the CSNs of cells sampled from case and control subjects to reveal intriguing patterns in gene co-expression.

1 INTRODUCTION

Single-cell RNA-sequencing (scRNA-seq) provides a high throughput profile of RNA expression for individual cells that reveals the heterogeneity across cell populations. Recent advances in computational methods enable cell-type classification, novel cell-type identification (38), and trajectory alignment (29); however, among these methods, less attention has been devoted to gene-gene association and transcriptional networks, which can shed light on many vital biological processes. Gene expression is known to be tightly regulated by networks of transcription factors and understanding these networks at a cellular level can identify differences in the inner workings of cells in normal and diseased tissues (10, 39, 27). Thus far, efforts to estimate co-expression networks from scRNA-seq have been limited in their success (11, 34) for several reasons, including technical challenges such as a sparsity of non-zero counts and a high level of noise (5), nonlinear relationships that are not easily captured by traditional measures, and heterogeneity in co-expression patterns across cells.

Efforts have been made to circumvent these challenges (20, 21, 39, 13, 27), but success has been limited. Many approaches aim to estimate a single co-expression network across all cell-types, but to gain access to the subtle differences in gene co-expression at the cellular level we require estimates of gene networks evaluated for each cell-type. Moreover, even if the aim is to estimate a cell-type specific network, traditional methods estimate a single co-expression relationship across the entire sample of cells of that type. With such an approach heterogeneity of co-expression across individual cells is erased, but this knowledge can provide valuable insights for downstream analysis.

To reveal co-expression networks within individual cells, we aim to construct cell-specific networks (CSNs) on the cellular level from scRNA-seq data. The idea of CSN was first proposed by Dai et al. (7) and Li et al. (19), wherein they generate individual transcriptional networks at a single-cell level for the first time. For each cell, they then record the degree sequence of its cell-specific network, to be used as input for clustering or low-dimensional embedding.

We propose a refinement (called locCSN for local Cell-Specific Networks) of the original construction method, which exhibits greater power to detect. Additionally, we demonstrate new downstream analyses that greatly expand the utility of CSNs for gaining insights into cell biology. Most notably, we retain the individual networks, rather than just their degree sequences, so that differences or developmental changes can be examined at the level of the gene-gene co-expressions. For instance, evolving gene communities can be estimated in developmental trajectories, or gene pairs with high differential co-expression can be identified when testing the networks for differences between subjects with different phenotypes.

The construction of CSN is based on the following assessment of local statistical dependency of a pair of genes: for a cell of a particular type, if the co-expression of the pair of genes is unusually high, relative to their distribution if the genes are assumed to be independent, then we infer a gene-gene relationship (Figure 1). This type of analysis produces a full gene-gene network for each cell, and the outcome for each cell is envisioned as a graph, with genes as nodes and edges depicting gene-gene dependencies. Although each individual network is estimated with considerable noise, in concert these CSNs can be used to perform a number of statistical analysis not previously possible.

CSN information is available to obtain several key elements that enhance our ability to model gene networks. First, averaging CSNs across cells within a category provides a clearer estimate of the gene network than competing methods, such as those based on Pearson's correlation, because of the sensitivity to nonlinear measurements, local dependencies and cellular heterogeneity. Second, having a sample of CSNs for each cell-type provides a valuable measure of the variability across cells that can be leveraged in powerful tests of network differences across cell groups. For example, comparisons between cells obtained from case and control subjects, or cells sampled from different spatial or temporal regimes. Finally the CSN approach facilitates followup analysis to discover the key genes driving the differences between networks. With this approach we can identify 'differential network (DN) genes', which do not differ in gene expression, but do differ substantially in terms of the co-expression network. These genes could play a key role in network regulation and hence differences that might help explain etiology of disease.

In this paper, we illustrate the value of CSN by applying locCSN to several scRNA-seq datasets, including developing fetal brains (26), and case and control samples from autism spectrum disorder (ASD) subjects (36). locCSN extracts cell level network information from these data which preserve cell-level network heterogeneity and highlights network differences between case and control samples shedding light on important functional differences.

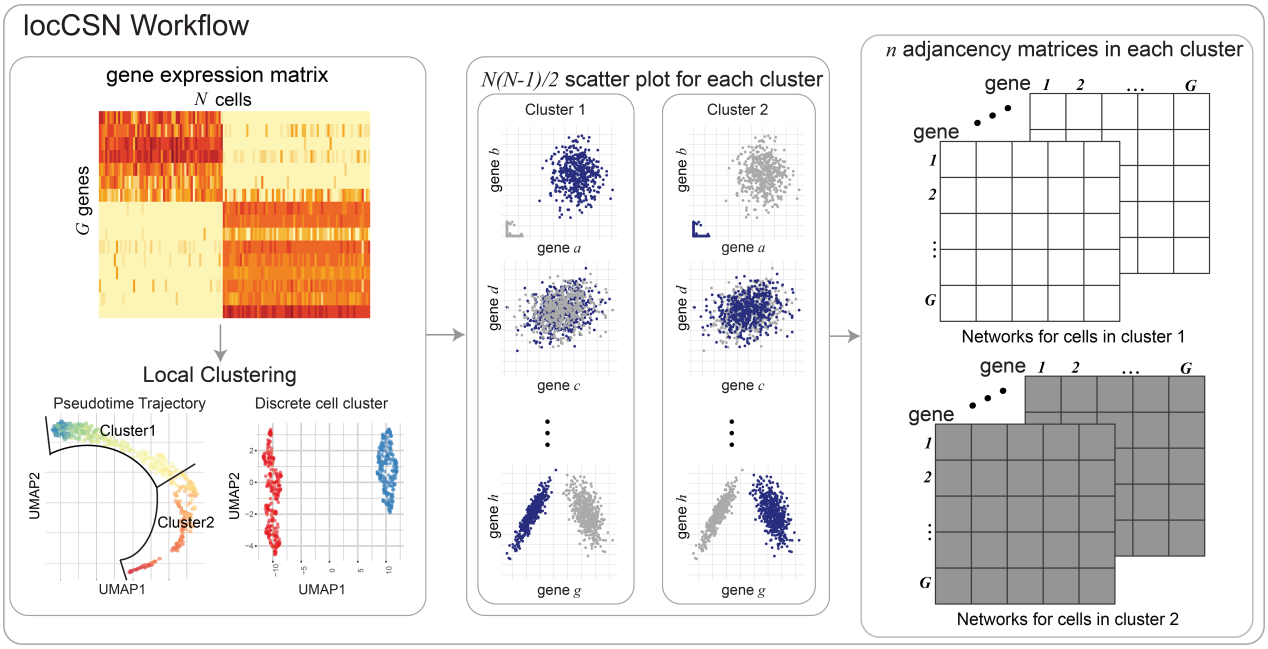


Figure 1: Workflow of locCSN. Starting with the gene expression matrix, we cluster homogeneous cells. For all cells in a cluster, we examine the co-expression of each pair of genes to determine if the joint expression is unusually high, relative to expectation, assuming the genes are independent. For each cell, we then construct the gene-gene network, based on the results of a test for local independence for each pair of genes.

2 MATERIALS AND METHODS

The algorithm for construction of CSN as originally published Dai et al. (7) (oCSN) is based on assessing local statistical dependency using all profiled expression of cells of all cell-types; however, oCSN can lead to false edges due to Simpson's paradox. For example, consider a pair of marker genes that display independent expression within the marked cell-type. oCSN will likely infer an edge due to the differential expression across cell-types. Another drawback of the original algorithm is the fixed number of cells used in the local independence test, which may significantly reduce detection power in some settings.

We propose a new local algorithm for CSN construction (locCSN) that overcomes these two drawbacks by restricting all calculations to a local neighborhood, and by allowing the bandwidth of the local independence test to vary by cell. Our local approach can be implemented in a variety of ways to avoid confounding, with the choice depending on available information. If cells have been reliably classified by cell-type, then the CSN calculations can be performed independently for each category, with extension to soft clustering weights if the cell-type labels are noisy. If uncertainty of cell-type arises because cells are developing along a smooth trajectory, then the cells can be windowed according to pseudotime (29), computing each CSN within its local bin. For each of these options, locCSN avoids Simpson's paradox by controlling for cell-type, and hence yields clearer co-expression patterns than oCSN and many traditional measures of correlation that ignore cell-type.

The recent paper (19) proposes a different modification of the oCSN method, in which the independence test is replaced by a conditional independence test. Under this method, an edge between two genes in a network represents dependence that cannot be explained by a third gene acting as a common driver for both. As such, the resulting network (termed Conditional CSN, or CCSN) is comparable to a partial correlation or conditional independence network, while oCSN and locCSN return networks that are more comparable to correlation networks. Both types of networks are useful for understanding cell behavior in different ways; for example, correlation networks are commonly used for detection of modules or community structure (18).

2.1 The locCSN method

To avoid confounding, we aim to construct CSNs using only the cells of a common type in the calculations. In settings where cell-type labels have been assigned, CSN construction can be applied separately to each cluster, as illustrated for a single cluster in Figure 1.

Given N cells and G genes, CSN construction takes as input the $N \times G$ matrix X of expression levels, and returns N matrices $A^{(1)}, \ldots, A^{(N)}$ of size $G \times G$, where $A^{(j)}$ is the estimated gene-gene relationships for cell j. Following Dai et al. (7), for a cell $j \in 1, \ldots, N$ and genes x and y in $1, \ldots, G$, let $D_x^{(j)}$ and $D_y^{(j)}$ denote one-dimensional bins for genes x and y centered at the expression levels for cell j, with widths w_x and w_y:

$$\begin{aligned} D_x^{(j)} &= \{i : |X_{ix} - X_{jx}| \leq w_x\} \text{ and} \\ D_y^{(j)} &= \{i : |X_{iy} - X_{jy}| \leq w_y\}, \end{aligned}$$

with cell counts $n_x^{(j)} = |D_x^{(j)}|$ and $n_y^{(j)} = |D_y^{(j)}|$. Let $D_{xy}^{(j)} = D_x^{(j)} \cap D_y^{(j)}$ denote the joint window centered at cell j with counts $n_{xy}^{(j)} = |D_{xy}^{(j)}|$.

For genes x and y in cell j, the estimated gene-gene relationship $A_{xy}^{(j)}$ is represented by a normalized test statistic, $A_{xy}^{(j)} = \rho_{xy}^{(j)} / \sigma_{xy}^{(j)}$, where $\rho_{xy}^{(j)}$ is a local test statistic for independence of genes x and y, comparing the joint distribution with the product of the marginals,

$$\rho_{xy}^{(j)} = \frac{n_{xy}^{(j)}}{N} - \frac{n_x^{(j)} n_y^{(j)}}{N^2}, \tag{1}$$

and the normalizing factor $\sigma_{xy}^{(j)}$ is the asymptotic standard deviation of $\rho_{xy}^{(j)}$ under the null hypothesis that genes x and y are independent, shown in Dai et al. (7) to equal

$${\sigma_{xy}^{(j)}}^2 = \frac{n_x^{(j)} n_y^{(j)} (N - n_x^{(j)})(N - n_y^{(j)})}{N^4 (N-1)}. \tag{2}$$

To motivate this construction, Dai et al. (7) give examples where the statistic succeeds in separating a mixture of cells, where genes x and y follow a dependent relationship in some cells but are independent in others (Supplementary Figure S2**a**). However, the choice of the window sizes w_x and w_y plays an important role in the performance of the algorithm. Dai et al. (7) proposed windows based on quantiles, but we find that this approach gives counter-intuitive results and low power when the joint distribution of genes x and y

is a simple correlated normal (Supplementary Note 1). In our algorithm, we instead choose the window size based on a local standard deviation, which we compute iteratively. Bins $D_x^{(j)}(0)$ and $D_y^{(j)}(0)$ are initialized to include a quantile range (as in Dai et al. (7)), and then iteratively we follow

$$w_x^{(j)}(t) = \text{St. Dev}\{X_{ix} : i \in D_y^{(j)}(t-1)\} \quad (3)$$

$$w_y^{(j)}(t) = \text{St. Dev}\{X_{iy} : i \in D_x^{(j)}(t-1)\} \quad (4)$$

$$D_x^{(j)}(t) = \{i : |X_{ix} - X_{jx}| \leq w_x^{(j)}(t+1)\} \quad (5)$$

$$D_y^{(j)}(t) = \{i : |X_{iy} - X_{jy}| \leq w_y^{(j)}(t+1)\} \quad (6)$$

for $t = 1, \ldots$ until convergence is achieved (Supplementary Figure S2**b**). We find that this approach gives more interpretable results for our simulated and data examples.

We remark that as a pre-processing step, we fix $A_{xy}^{(j)} = 0$ if the expression of either gene is zero for that cell.

2.2 locCSN for developing and heterogeneous cells

To construct CSNs from a mixture of cells of different cell-types, when distinct cell-type labels are not available, or when cell-types are not clearly delineated, modifications are required. One option is to utilize soft clustering to remove the confounding effect that arises with poorly classified cells. Alternatively, when cells are more naturally ordered along a developmental trajectory, then we suggest binning cells based on pseudotime to obtain local neighborhoods and then compute CSN within the local neighborhoods.

Soft clustering Soft clustering can directly remove the confounding effect that arises with poorly classified cells. In our method locCSN, we use SOUP (41) to estimate the soft clustering probabilities and treat them as inputs. Let K denote the number of clusters, and for each cell j let $s_j = (s_{j1}, \ldots, s_{jK})$ denote its soft clustering weights, normalized so that $\sum_k s_{jk} = 1$. For each cell j and pair of genes x and y, locCSN estimates the gene-gene relationship for that cell by a weighted sum of independence test statistics across all K clusters,

$$A_{xy}^{(j)} = \sum_{k=1}^{K} s_{jk} \frac{\rho_{xyk}^{(j)}}{\sigma_{xyk}^{(j)}},$$

where $\rho_{xyk}^{(j)}$ and $\sigma_{xyk}^{(j)}$ are computed using analogously to $\rho_{xy}^{(j)}$ and $\sigma_{xy}^{(j)}$ as given by (1) and (2), but with each cell i weighted by its soft clustering weight s_{ik}. That is, the bin cell counts such as $n_x^{(j)}$ appearing in (1) and (2) are replaced by the bin weight sums, e.g., $n_x^{(j)} = \sum_{i \in D_x^{(j)}} s_{ik}$, with N replaced by $\sum_i s_{ik}$. To find the bins $D_x^{(j)}$ and $D_y^{(j)}$ for a particular cluster k, the initial bin widths $w_x^{(j)}(0)$ and $w_y^{(j)}(0)$ are taken to be a $q\%$ weighted quantile window,

$$w_\ell^{(j)}(0) = \text{argmin}_d \text{ such that}$$
$$\sum_i s_{ik} 1\{|X_{i\ell} - X_{j\ell}| \leq d\} \geq q \cdot \sum_i s_{ik},$$

for $\ell = x$ or y; then the iterates (3)-(6) are followed as before, except that the standard deviations in (3)-(4) are computed with respect to the weights $\{s_{ik}\}_{i=1}^N$.

Trajectory analysis It is especially challenging to cluster cells by cell-type when they are developing along a smooth trajectory. In this setting, using trajectory analysis to obtain estimates of pseudotime, cells can be sorted by developmental stage. After which, locCSN can be performed on bins consisting of cells with similar pseudotimes. The binning method is convenient for downstream estimates of CSNs by developmental epoch. Alternatively, locCSN can be applied using the m-nearest neighbors based on pseudotime coordinates.

2.3 Additional Methods

Metacell: Reduce sparsity of expressions For single-cell and single-nuclei RNA-seq datasets, the expression can be very sparse. In this setting a direct application of the CSN algorithm fails to discover network structure. It can be advantageous to cluster the data before performing downstream analysis (17), hence, we apply the Metacell algorithm (3) before constructing CSNs. Metacell partitions cells into metacells,

defined as disjoint clusters of homogenous profiles. After applying Metacell to pre-labeled cells, we further divide metacells with multiple cell-types or subtypes into pure-cell-type metacells. Expression of a metacell is defined as the mean of the cells in the cluster, which alleviates the problem of having zero expression for many genes per cell. The metacells are then treated as cells for the purpose of constructing metacell-specific-networks. In this paper, for convenience, CSN refers to either a cell-specific-network or metacell-specific-network.

Windowed analysis Within a given cell-type, cells may exhibit heterogeneity that is poorly delineated into subclusters. In such cases, it may be prudent to even further localize the CSN computations. As a simple heuristic for doing so, for each cell (or metacell) we can construct its CSN using only a fixed number or proportion of the cells in its cell-type, corresponding to that cell's closest neighbors after dimension reduction such as UMAP.

Network estimation For each cell, locCSN produces a test statistics for each pair of genes, A_{xy}, to evaluate pairwise gene-gene independence in the neighborhood of that cell's expression. By thresholding these tests we obtain a 0-1 adjacency matrix for each individual cell. Since A_{xy} is approximately normally distributed, a natural threshold is the standard normal quantile, Z_t for t chosen to correspond with a natural choice for significance testing, such as 0.95 or 0.99. An estimate of the network for a given cell-type or cluster of cells is obtained by averaging the adjacency matrices, which we denote by $\bar{A}$.

Trajectory-based community detection The PiSCES algorithm (21) is designed to identify cluster structure that varies smoothly over adjacent developmental periods. Following CSN construction for a smooth trajectory, where the cells have been binned by similar psuedotimes, we apply the PiSCES algorithm using the average CSN of each bin as input. The time-varying community structure can then be visualized using Sankey plots.

Two sample testing Given i.i.d. samples of expression levels, the computed CSNs for a set of cells are exchangeable, and hence permutation testing can be used to test for differences in CSN distribution. For this purpose, we suggest two types of tests: first, an omnibus test for generic differences, and second, a targeted test, aimed at identifying high leverage genes that drive the difference.

2.3.1 DISTp: Test CSN differences between groups

To test for differences in the distribution of the two classes, we can use a nonparametric distance-based test statistic (DISTp) taken from (22), coupled with a permutation test. This test is highly sensitive to differences in network structure, but a rejection merely indicates that the networks are significantly different.

Each cell's adjacency matrix can be represented as a vector, resulting in n_1 sample vectors from class 1 and n_2 sample vectors from class 2. Let $V_1^{(1)}, \ldots, V_{n_1}^{(1)}$ denote the vectorized adjacency matrices from class 1 and $V_1^{(2)}, \ldots, V_{n_2}^{(2)}$ denote the same from class 2. The test statistic Q is a scaled α-norm divergence measurement, with $\alpha \in (0, 2)$ recommended (22), and is given by

$$\begin{aligned} Q(\mathbf{V}^{(1)}, \mathbf{V}^{(2)}; \alpha) \quad = \quad & \frac{2}{n_1+n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} |V_i^{(1)} - V_j^{(2)}|^\alpha - \frac{n_1 n_2}{n_1+n_2} \binom{n_1}{2}^{-1} \sum_{1 \le i < i' \le n_1} |V_i^{(1)} - V_{i'}^{(1)}|^\alpha \\ & - \frac{n_1 n_2}{n_1+n_2} \binom{n_2}{2}^{-1} \sum_{1 \le j < j' \le n_2} |V_j^{(2)} - V_{j'}^{(2)}|^\alpha, \end{aligned}$$

with p-value calculated by permutation test.

2.3.2 sLED: Identify leverage genes for co-expression differences between groups

To gain further insights into the network differences, we can utilize the sparse-Leading-Eigenvalue-Driven (sLED) test (40). sLED is designed to detect signal attributable to a modest number of genes in the high dimensional setting encountered for studies of transcription. To emphasize the contrast with differentially expressed genes, we call these differential network (DN) genes.

This test relies on the same principles as Sparse Principal Component Analysis (SPCA), and was originally proposed for the difference in the Pearson's correlation matrices of the two classes (sLED-Pearson). Here we instead propose using the difference in the average CSN as the test input (sLED-CSN). Given n CSN

adjacency matrices from class 1 and m from class 2, denoted by $A_1^{(1)}, \ldots, A_n^{(1)}$ and $A_1^{(2)}, \ldots, A_m^{(2)}$, let D denote the difference between the average CSN for each class, so that $D = \bar{A}^{(1)} - \bar{A}^{(2)}$. Then D can be used as the input to sLED, in which case the test statistic is computed from the spectrum of D. Additionally, the test also identifies a small cluster of leverage genes corresponding to the non-zero entries of the sparse leading eigenvector. These are candidate genes that have altered co-expression structure between the two groups. As with DISTp, the p-value of the test statistic is determined by permuting samples among classes.

2.4 Datasets for analysis

Chu-type dataset (6) includes 1018 cells and seven cell-types. This dataset contained the cells of human embryonic stem cell-derived lineage-specific progenitors. The cell-types including H1 embryonic stem cells, H9 embryonic stem cells, human foreskin fibroblasts (HFF), neuronal progenitor cells (NPC), definitive endoderm cells (DEC), endothelial cells (EC) and trophoblast-like cells (TB) were identified by fluorescence-activated cell sorting (FACS) with their respective markers. 9600 genes are obtained per cell on average.

Developing cortex atlas dataset (26) includes cells from mid-gestational human cortex. These data are derived from ∼40,000 cells from germinal zones (ventricular zone [VZ], subventricular zone [SVZ]), developing cortex (subplate [SP] and cortical plate [CP]) separated before single cell isolation.

Autism spectrum disorder (ASD) brain dataset (36) includes snRNA-seq data from an ASD study, which collected 105 thousand nuclei from cortical samples taken from 22 ASD and 19 control samples from subjects between 4 and 22 years old. Samples were matched for age, sex, RNA integrity number, and postmortem interval.

3 RESULTS

Our results are summarised as follows. First, simulated and real data examples are used to illustrate differences between locCSN and oCSN. Second, we apply time-varying community detection to CSNs taken from two overlapping developmental trajectories in the developing cortex atlas dataset, and identify differences in their community evolution. Finally, we show that CSNs enable better separation of case and control populations in the ASD brain dataset, by identifying genes that have differing co-expression (but similar expression levels) between the two groups.

3.1 Illustrative Examples

To illustrate the importance of local calculations we compare locCSN and oCSN for synthetic data from ESCO (35). The cells are sampled from two populations, one for which a set of genes exhibit pairwise correlation and another for which the genes are independent (Figure 2**a-c**). When calculated for this set of genes, ideally the test statistics for each cell will distinguish the pattern of correlation present in one population (Group2) not in the other (Group1) (Figure 2**c**). The desired pattern is achieved by locCSN, but oCSN produces false signals of correlation for many cells in Group1 (Figure 2**d** and **e**). Mixing the two populations leads to false positives in Group1 suggesting that the false positives result from Simpson's paradox.

Next we simulate data from ESCO, with 10000 genes and 2500 cells. We focus on 125 housekeeping genes that are not correlated with each other. After down-sampling the read counts to weaken the signal, we compare BigSCale (13) and locCSN with metacells. With no correlation between genes, we should not detect connections between genes. From 2500 cells, we constructed 158 metacells and on average, there are 15 cells per metacell. The heatmaps of Pearson's correlation of down-sampled and true read counts are shown in the first two panels of Figure 2**f**. The BigSCale correlation shows false positives between genes when there are no connection between genes. By contrast, average CSN with metacells shows no connection between genes.

To demonstrate and evaluate how soft clustering impacts the performance of CSN test statistics, we use the Chutype dataset (6), coupled with SOUP (41), a semi-soft clustering method for scRNA-seq data, which provides cluster membership probabilities. The Chutype dataset includes 7 cell-types (Figure 3**a**). Marker genes corresponding to developmental lineage are provided by the authors and a heatmap of gene expression reveals that a subset of genes mark cell-types DEC and NPC fairly well (Figure 3**b**). We analyze these two cell-types, which contain 138 cells and 173 cells, respectively. Although the cell-types are fairly distinct in UMAP, the soft cluster probabilities are not identically 1 or 0 (Figure 3**c**), indicating some uncertainty about group membership. The absolute Pearson's correlation for lineage marker genes, computed within DEC and NPC cell-types, do not show a clear pattern; specifically, the correlation does not delineate the expected block

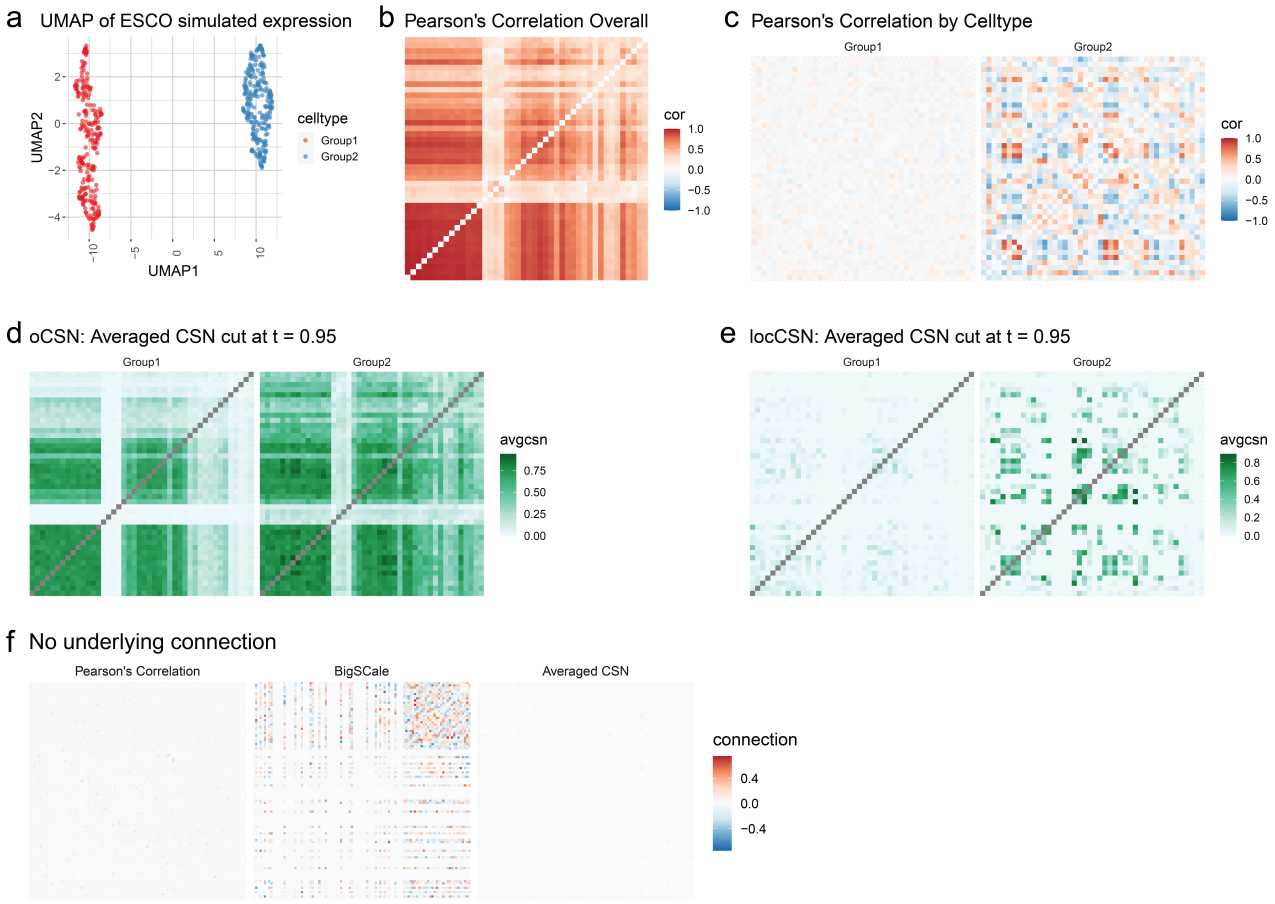


Figure 2: Network estimation for two simulated cell-groups. (**a**) UMAP from ESCO simulated gene expression. (**b**) Heatmap of Pearson's correlations of genes, calculated ignoring cell-types. (c) Heatmaps of Pearson's correlations of genes, calculated independently for each cell-type. (**d**) Heatmap of averaged oCSN within cell type, calculated ignoring cell-types. (**e**) Heatmaps of averaged locCSN within cell-type, calculated independently for each cell-type. (**f**) For a dataset simulated using ESCO with no correlation between any genes, heatmaps of Pearson's correlation, estimated BigSCale network and averaged locCSN.

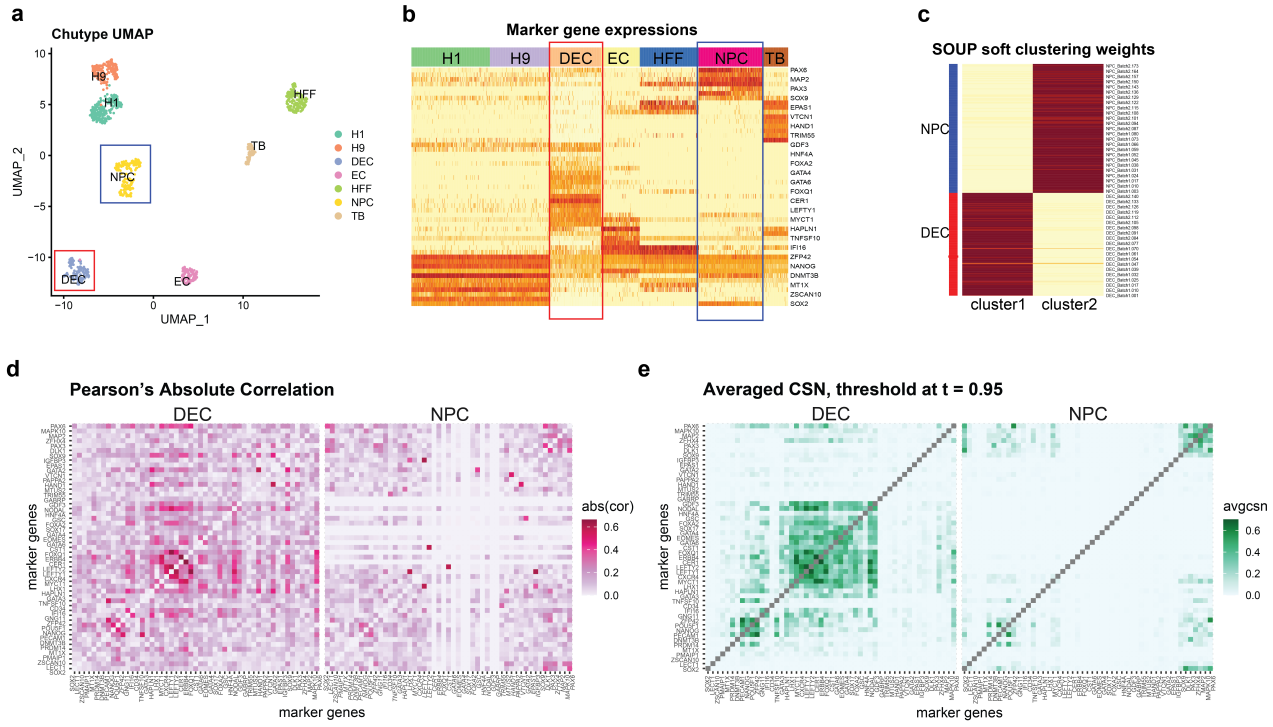


Figure 3: Estimated networks for Chutype cells. (**a**) UMAP of Chutype cells, colored by cell-types. The red and blue boxes indicate selected cell-type: DEC and NPC; (**b**) Heatmap of gene expression for 57 developmental genes for 7 cell-types. High to low expression corresponds to red to light yellow in heatmap. 7 cell-types are color coded by the band on the top of the heatmap. The red and blue boxes indicate DEC and NPC cell-types; (**c**) Heatmap of estimated soft clustering weights from SOUP for DEC and NPC cell-types. The band on the left of the heatmap shows the NPC and DEC cell-types; (**d**) Heatmaps of absolute Pearson's correlation for DEC and NPC cell-type, calculated independently within cell-types; (**e**) Heatmaps of averaged locCSN for DEC and NPC cell-types, thresholded at Gaussian distribution 95% quantile.

structure for marker genes for DEC and NPC cells (Figure 3**d**). By contrast, averaged locCSNs, thresholded at normal distribution 95% quantile, preserve the gene block structure and emphasizes the differences between cell-types (Figure 3**e**). We conclude that CSN with soft clustering works well on distinct cell-types. The averaged CSN preserves gene blocks, distinguishes between cell-groups, and depicts a clearer co-expression pattern than Pearson's correlation. The averaged CSNs calculated from oCSN ignore co-expression between genes, especially the dense block for NPC at the upper right corner (Supplementary Figure S3).

3.2 CSN analysis of developmental trajectories

To illustrate the the application of CSN to developing cells we feature the developmental trajectory of excitatory neuron cells in the Developing cortex atlas dataset: Radial glia and Progenitors (P), Intermediate Progenitors (IP), Maturing Excitatory (ExN), Migrating Excitatory (ExM), Excitatory Upper-layer entriched (ExM-U), and Excitatory Deep layer (ExDp) from the brain cortex atlas dataset (26), with cell-type labels determined by the authors. Using Slingshot (31), we estimate the developmental path consists of two trajectories: one ending in upper layer (U-curve) and other in deep layer (D-curve) excitatory neurons (Figure 4**a** and **b**).

To circumvent challenges due to sparse counts, which are prevalent in these data, we pool similar cells within cell-type and form metacells (3), each containing approximately 20 cells. Each metacell's expression, pseudotime and UMAP coordinate is computed as the average over pooled cells. Metacells are assigned to trajectories based on proximity using Slingshot (31). At the early stages of development the curves are nearly overlapping and metacells are not differentiated by curve, but as the cells develop, metacells can be assigned to a distinct curve (Supplementary Table S3-S4). Both curves start at the P cell class with full overlap, but moving along ExN and ExM cell classes, there is progressively less overlap between the two curves until they bifurcate, with D-curve and U-curve culminating in purely ExDp and ExM-U metacells, respectively. Next, within each cell-type, we generate bins based on the metacell pseudotime values. This effectively creates bins of approximately 800 metacells containing fairly homogeneous cells.

For each metacell, we use locCSN to compute gene networks. For illustration we wish to focus on a restricted gene list. We choose ASD risk genes because the development of excitatory neurons is of interest in ASD research (26, 24, 30). Specifically CSNs were computed for 444 genes chosen by intersecting the expressed genes in the metacells with a list of 992 ASD-associated genes gleaned from the SFARI database (classes S, 1 and 2) (4). Our objective is to map the formation of gene-clusters over developmental epochs, as cells develop into upper or deep layer excitatory neurons. We apply PiSCES (21) to the average CSNs per bin, to find time-varying gene-community structure in the D-curve and U-curve trajectories, and use Sankey plots to visualize the evolution of gene-communities (Figure 4**c** and **d**). As expected the gene-communities are nearly identical for the two curves for the first four pseudotime bins, which share a large fraction of metacells, and the results progressively differentiate as the overlap in metacells diminish. Most notably, the gene-community associated with cluster 4 (purple) for U-curve splits and merges into the other gene-communities at ExN_pt3, while it persists for D-curve. For both curves, cluster 1 (red) contains genes that are more highly connected than the other loosely connected gene-communities (Supplementary Figure S5). For all of the remaining identified gene-clusters the correlation is extremely weak in the early stages and becomes more apparent in the final developmental bin.

We study the membership in gene-clusters when they are relatively stable without major splitting and merging, i.e. D-curve from ExN_pt1 to ExDp and U-curve from ExN_pt3 to ExM-U. We refer to gene-clusters in D-curve as cluster 1-4 and U-curve as cluster 1-3, where labels run from from top to bottom of the display. The numbers of genes in each stable cluster and the overlap between clusters from two curves are shown in the Table 1(Supplementary Table S11). For example, almost all of the genes from the dense cluster for D-curve are included in the corresponding dense cluster for U-curve. While the cluster membership is fairly stable along each curve, the strength of correlation changes over time (Supplementary Figure S5).

We also consider the averaged metacell gene expression across pseudotime for each gene-cluster (Figure 4**e** and **f**). Gene expression is relatively stable across cellular development. The dense gene-cluster, cluster 1 in both curves, has high expressions while the loose gene-clusters have relatively low expression. Yet, even though all genes outside of the dense cluster have lower expression levels, using locCSN, combined with PisCES, we are able to detect subtle correlation and partition genes into gene-communities. It is worth noting that none of these communities are identified by WGCNA (18) (Supplementary Figure S4), which relies on Pearson's correlation, but could also be implemented using average CSN matrices for potentially better performance.

To understand the function of gene communities, we check Gene Ontology (GO) terms (2) (Supplementary Note 5) for the seven gene communities (Supplementary Figure S6, Supplementary Table S13). Clusters 1-3 show similar results for both curves. GO term treemaps for the dense gene community (cluster 1) are

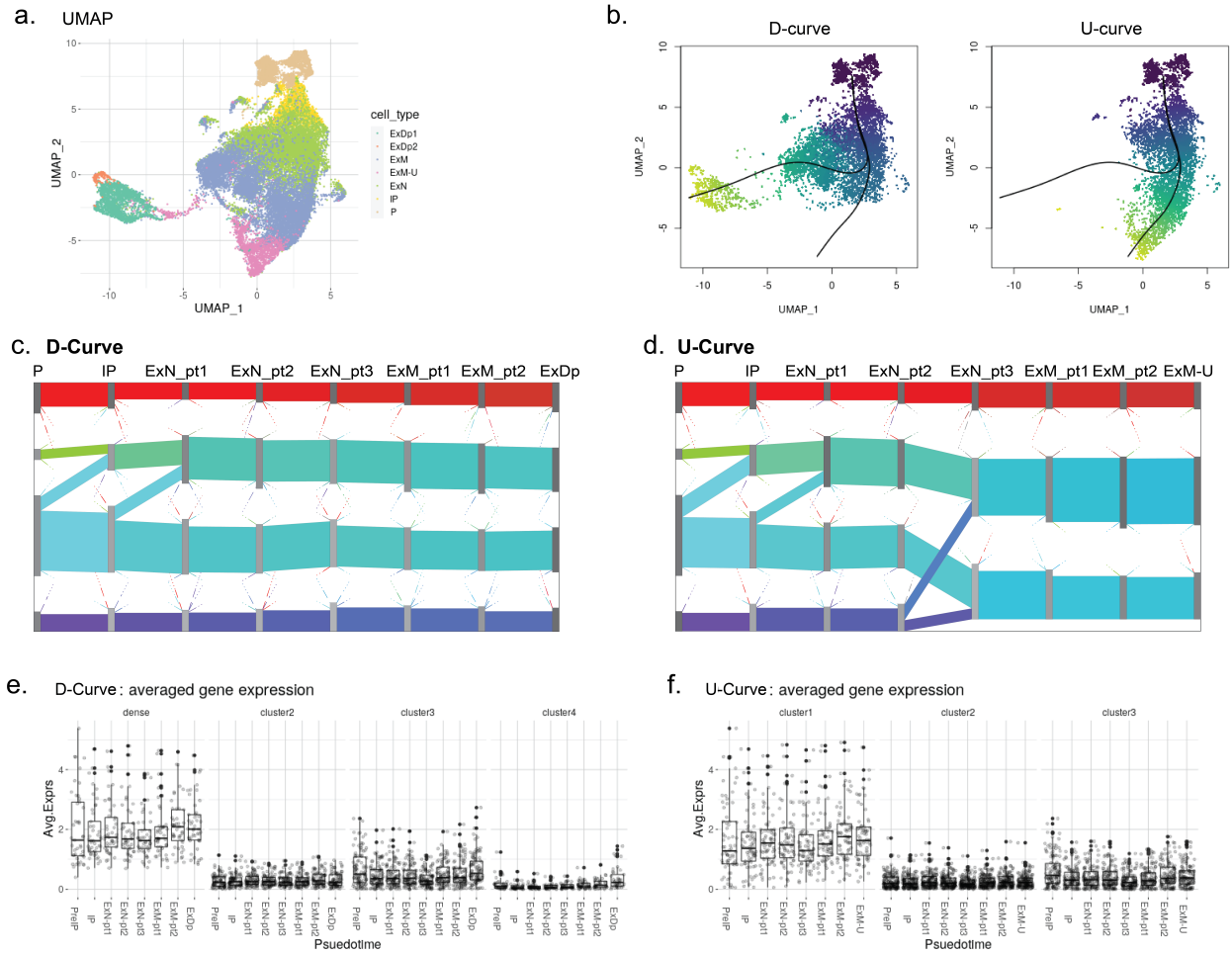


Figure 4: Development of networks in human fetal brain cells. (**a**) UMAP of human fetal brain single-cell expression from 7 cell-types involved in development of excitatory neuron cells, (**b**) with developmental trajectories superimposed. The UMAP plots indicate metacells, whose coordinates are the average UMAP cellular locations. The two principal curves generated by Slingshot (31). Colors are determined by the metacell's pseudotime for the two curves, which are calculated as the average pseudotime over all cells in the metacell. The left and right panel shows the metacells assignment based on pseudotime to D-curve and U-curve, respectively. (**c,d**) Sankey plots of averaged CSN for 7 bins in D-curve and U-curve. Gene flows are shown as the colored band connecting two adjacent pseudotime bins for (**c**) D-curve and (**d**) U-curve. (**e**) and (**f**) depict boxplots of averaged metacell gene expression for 7 pseudotime bins in the final 4 and 3 clusters for the D- and U-curves, respectively. The x-axis shows pseudotime bins and y-axis shows the averaged expression. Three clusters for U-curve.

		D-Curve				
		cluster 1	cluster 2	cluster 3	cluster 4	Total
U-Curve	cluster 1	43*	0	3	0	70
	cluster 2	0	67*	7	20	160
	cluster 3	0	4	82*	12	127
	Total	45	86	103	41	

Table 1: Number of genes in gene communities of D-curves and U-curve. The rows indicate 3 gene clusters from U-curve and the columns show 4 gene clusters from D-curve. The "total" row and column show the total number of genes in each gene cluster. The intersect between U-curve gene clusters and D-curve gene clusters are shown in each row and clusters. The sum of number of overlapped genes does not sum to the total number of genes. * indicates two clusters in D-Curve and U-Curve match with each other.

involved in the process of mitosis which is critical during the fetal stage. Cluster 2 is enriched for chromatin organization, which is critical for gene expression regulation. The most loosely structured set of genes, cluster 3 shows no GO enrichment. Of greatest interest is cluster 4, which is ultimately restricted to D-curve and enriched for synaptic organization, suggesting that these neurons are more mature. This discovery fits with the biological process of neural migration which naturally proliferates the deep layers earlier than the upper layers.

3.3 CSN analysis of brain cells from ASD subjects

We analyzed single-nuclei RNA-seq (snRNA-seq) data from an ASD study of cortical nuclei assessing 17 cell clusters: fibrous astrocytes (AST-FB), protoplasic astrocytes (AST-PP), endothelial (End), parvalbumin interneurons (IN-PV), somatostatin interneurons (IN-SST), SV2C expressing interneurons (IN-SV2C), VIP expressing interneurons (IN-VIP), upper-layer excitatory neurons (L2/3), layer 4 excitatory neurons (L4), deep layer cortico-subcortical excitatory projection neurons (L5/6), Deep-layer cortico-cortical excitatory projection neurons (L5/6-CC), microglia (Mic), immature neurons (Neu-mat), neurogranin expressing neurons I (Neu-NRGN-I), neurogranin expressing neurons II (Neu-NRGN-II), oligodendrocytes (Oli) and oligodendrocyte precursor cells (OPC) (36). For analysis, we merge some celltypes together as cell-groups depending on whether the cell-types are distinct in tSNE plot (Figure 5**a**, Supplementary Note 4, Supplementary Figure S7). For instance, the AST* cell-types are merged into one cell-group, while the L* cell clusters are distinct and analyzed individually (Figure 5**b**). To circumvent challenges due to sparse counts, which are especially prevalent in single-nuclei RNA-seq data, we cluster similar cells and form metacells (3), which contains around 20 cells each. To avoid batch effects, metacells are created within a sample and cell-type. (Supplementary Table S8)

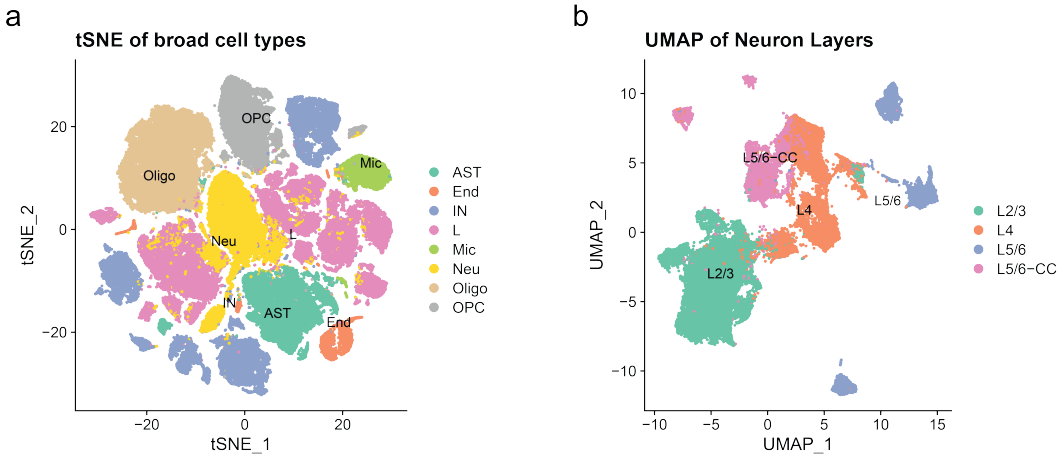


Figure 5: Dimension reduction of brain cells. (**a**) tSNE plot of all brain cells colored by broad cell-types. (**b**) UMAP plot of Neuron Layers(L) cells colored by 4 cell clusters.

For each metacell within a cell-group, we construct CSNs using the nearby 100 metacells from UMAP plot of the combined ASD and control cells (Supplementary Figure S8). CSNs are computed only for genes implicated in ASD in the literature. We obtain the gene list by intersecting the 992 measured SFARI genes (4) with expressed genes of ASD case and control dataset and 942 genes remain. After CSN construction, we compare the differences between control and ASD groups by sLED-CSN, sLED-Pearson and DISTp for each cell-type.

The sLED-Pearson test detects no significant differences between control and ASD groups for any cell-groups after adjusting for multiple testing, while the sLED-CSN and DISTp tests detect differences in 10 and 6 of the 13 cell-groups tested, respectively (Table 2, Supplementary Table S9). We illustrate our results using excitatory neurons layers: L2/3, L4, L5/6 and L5/6-CC. Leverage genes identified by sLED-CSN show striking differences in their averaged CSNs (Figure 6**a-d**, Supplementary Figure S9) and also their visualized networks (Figure 7). Comparing across layers, we find minimal overlap in leverage genes (Figure 8**a**). Finally, removal of the leverage genes reduces the number of detected differences from 10 to 2, suggesting that they explain much of the difference between the groups (Supplementary Table S10).

We performed a GO enrichment analysis with leverage genes (Table 2, Supplementary Table S12) identified by SLED-CSN for cell-groups that are significantly different. Among 10 significant cell-groups, only L5/6

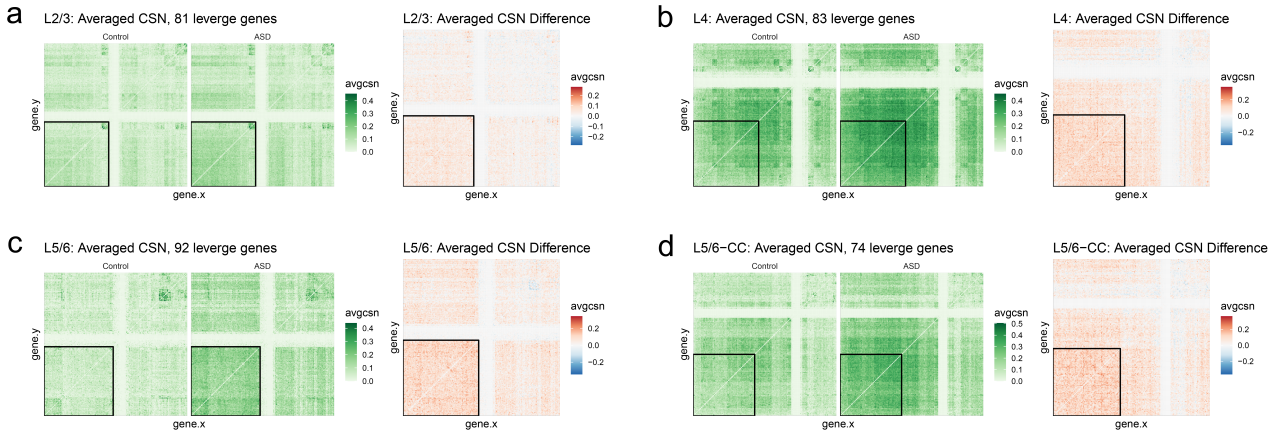


Figure 6: Heatmaps of average CSNs and the average difference (ASD minus Control). Heatmaps display sLED-CSN selected leverage genes for each cell-group, contrasted with an additional 100 randomly selected genes from the 942 ASD genes. The black squares delineate the leverage genes for each cell-group: (**a**) L2/3; (**b**) L4; (c) L5/6 ;and (**d**) L5/6-CC.

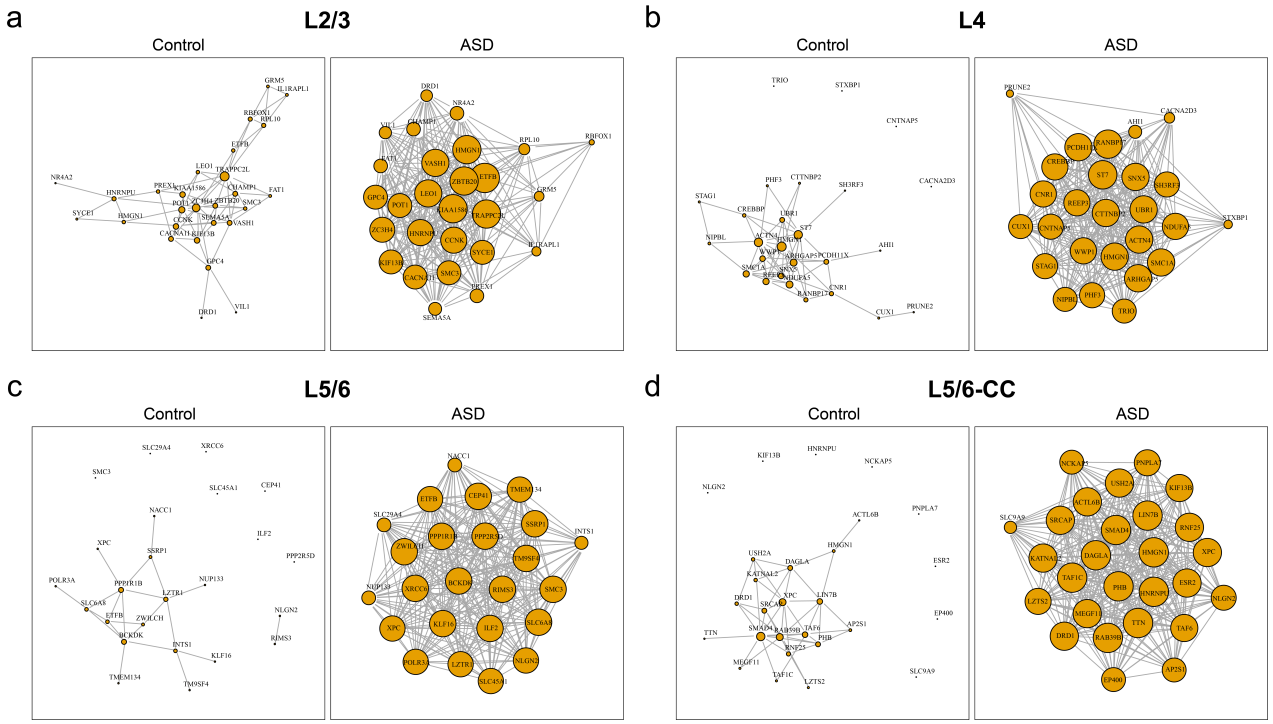
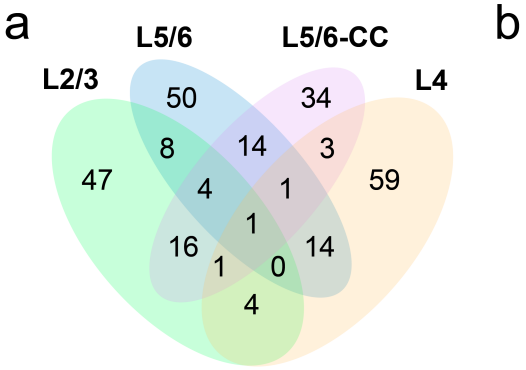


Figure 7: Gene networks for leverage genes in the excitatory neuron layers. The networks are generated from averaged CSN of control and ASD groups. Vertices are high leverage genes, defined as the top loading genes that explain 90% of the signal. Edges are determined by thresholding averaged CSNs and the size of the vertex corresponds to the degree of the vertex. (**a**) L2/3; (**b**) L4; (**c**) L5/6; (**d**) L5/6-CC.



b

Cell group	SFARI DE genes	DN genes
AST	12	26
IN-VIP,SV2C	18	26
IN-SST,PV	11	26
L2/3	16	27
L4	17	26
L5/6	1	24
L5/6-CC	2	27
Neu-NRGN	5	26
Oli	0	28
OPC	1	27

Figure 8: Contrasting differentially expressed and leverage genes for ASD vs. Controls. (**a**) Venn diagram of all leverage genes for neuron layer cell-types. (**b**) Numbers of ASD differentially expressed (DE) genes and differential network (DN) genes, which have no overlap.

and L5/6-CC have significant GO terms (Supplementary Table S14). For L5/6, we observe enrichment of regulation of phosphate activity and dephosphorylation, both of which map onto protein function and cell signaling. For L5/6-CC, we observe enrichment of chromosome organization, which impacts gene expression.

P-values	L2/3	L4	L5/6	L5/6-CC
sLED-CSN	0.001*	0.001*	0.001*	0.001*
sLED-Pearson	0.089	0.861	0.564	0.348
DISTp	0.001*	0.039	0.001*	0.001*

Table 2: P-values for comparing gene expressions and CSNs for control and ASD groups by cell-group. Comparing gene expressions and CSNs for control and ASD groups, p-values are calculated by sLED-CSN, sLED-Pearson and DISTp for 4 excitatory neuron layers:L2/3, L4, L5/6 and L5/6-CC. * indicates significant differences between two groups after adjusting for multiple testing with 13 cell-types.

For the 10 significant cell-groups detected by sLED-CSN, we investigate the overlap between the differential expressed (DE) SFARI genes provided in (36) and the sLED-CSN high leverage genes (defined as the top loading genes that explain 90% of the signal, see Supplementary Table S11). We call these high leverage genes "differential network (DN) genes", to indicate that they strongly differ between case and control in their co-expression and network structure, but not necessarily in expression. There is surprisingly no overlap (Figure 8**b**), suggesting that most leverage genes do not show practically significant differences in expression level between control and ASD groups (Supplementary Table S11); similar conclusions can also be drawn by directly inspecting the marginal expression levels (Supplementary Figure S10).

4 DISCUSSION

Single-cell gene co-expression networks can yield critical insights into biological processes, but they are challenging to estimate due to cellular heterogeneity and sparsity of transcript counts. Our proposed solution, locCSN, utilizes a novel approach to produce a gene network for every cell, called CSN, using a method inspired by Dai et al. (7). We demonstrate in practice that locCSN performs better than the original oCSN algorithm by avoiding confounding due to cell-type heterogeneity and by using tuning parameters that yield a more powerful test for dependencies. While each CSN is estimated with considerable noise, averaging CSNs over homogeneous cells can provide stable estimates of network structure and this provides insights into how these networks vary by cell-type and over developmental epochs. Due to the nonparametric approach, averaged CSN networks provide better estimates of gene block structures than traditional measures. As a bonus, by exploiting the repeated estimates of network structure, the approach can be paired with powerful methods for testing network differences between groups of cells.

Understanding cell-type specific gene networks can contribute substantially to our understanding of how biological processes evolve over time in normal cells and how they are impacted by disease and disorders (23, 40, 12, 26). Just as we can test for differential gene expression, we can test for differential coexpression and aim to detect high leverage genes that drive the differences in coexpression. Small changes in gene

expression can lead to substantial changes in networks structure, ultimately with large biological effect. These differential network genes were called 'dark' genes by (7) because we fail to detect them with traditional differential expression and yet can detect them with CSN analysis. For example, we found that most genes that leverage significant differences of connection between ASD and control groups are missed when simply comparing their expression levels.

CRISPR-Cas technologies (8, 9, 15) provides researchers with tools to introduce and assess the effects of many genetic perturbations. For example, Jin et al. (16) used the Perturb-Seq method (9) to introduce dozens of ASD risk genes to developing mice brains and then assessed the impact on the single-cell transcriptome. Because CSN estimates the network of each cell it is a natural tool for analysis of such perturbation experiments, which target individual cells with distinct perturbations. CSN analysis provides a powerful tool that will facilitate the analysis of how such perturbations impact the network structure of a cell.

Constructing gene networks using the CSN approach is computationally intensive because we compute a test statistics for each pair of genes and each cell. Thus the computational complexity of constructing CSNs with locCSN increases quadratically with the number of genes and linearly with the number cells. Moreover, for each cell, locCSN depends on a window size based on the local standard deviation for each cell and each pair of genes. For sparse datasets, we can reduce the computational complexity substantially by replacing cells with metacells (3). We also provide an approximate CSN calculation by partitioning the outcome space for each pair of genes into a grid. Cells that fall into the same grid yield the same test statistic. With these approximations CSN can be readily applied to very large datasets.

It is often natural to reduce the genes under investigation by CSN to a meaningful subset, such as genes previously implicated in genetic risk, genes mapped to critical pathways, or highly variable genes. Restricting the investigation to a subset of genes greatly reduces the computational complexity of CSN analysis, but more importantly, it can reveal more scientifically interpretable results. By focusing on hundreds of documented ASD risk genes, we were able to identify intriguing functional network structure in developmental trajectories and changes in network structure between ASD and control subjects. In the literature several papers have implicated particular cell-types, especially neurons and subtypes of neurons, in ASD risk (1, 25, 26, 30, 36, 37), but no consensus has been reached. Here we found that gene network structure differed subtly between the developmental trajectories of fetal brain cells for upper layer and deep layer excitatory neurons. Specifically, while both trajectories revealed clustering of genes involved in gene expression regulation, only the deep layer trajectory showed clustering of synaptic genes. In the analysis of leverage genes, which drive differences between gene networks in ASD and control brain cells, we found an enrichment of phosphorylation genes in deep layer cortico-subcortical excitatory projection neurons (L5/6), but the cortico-cortical excitatory projection neurons (L5/6-CC) were primarily enriched for chromosome organization genes. Identifying differences in gene networks, both over developmental epochs and between phenotypes can shed light on the genetic etiology of diseases and brain disorders.

5 DATA AVAILABILITY

The datasets supporting the conclusions of this article are listed in Supplementary Material.

6 Software

The source code is available in this package (http://github.com/xuranw/locCSN).

7 SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

8 ACKNOWLEDGEMENTS

The authors are grateful to Kevin Lin and Jinjin Tian for helpful comments.

9 FUNDING

This work was supported in part by National Institute of Mental Health (NIMH) grants R01MH123184 and R37MH057881.

9.0.1 Conflict of interest statement.

None declared.

References

[1] Anahita Amiri, Gianfilippo Coppola, Soraya Scuderi, Feinan Wu, Tanmoy Roychowdhury, Fuchen Liu, Sirisha Pochareddy, Yurae Shin, Alexias Safi, Lingyun Song, Ying Zhu, André M M Sousa, PsychENCODE Consortium, Mark Gerstein, Gregory E Crawford, Nenad Sestan, Alexej Abyzov, and Flora M Vaccarino. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science*, 362(6420), 12 2018.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. Metacell: analysis of single-cell rna-seq data using k-nn graph partitions. *Genome biology*, 20(1):1–19, 2019.

[4] Saumyendra N Basu, Ravi Kollu, and Sharmila Banerjee-Basu. Autdb: a gene reference resource for autism research. *Nucleic Acids Res*, 37(Database issue):D832–6, Jan 2009.

[5] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1):232, 06 2018.

[6] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.

[7] Hao Dai, Lin Li, Tao Zeng, and Luonan Chen. Cell-specific network constructed by single-cell rna sequencing data. *Nucleic acids research*, 47(11):e62–e62, 2019.

[8] Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled crispr screening with single-cell transcriptome readout. *Nat Methods*, 14(3):297–301, 03 2017.

[9] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, Dec 2016.

[10] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front Cell Dev Biol*, 2:38, 2014.

[11] Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: A pipeline for single-cell rna-seq profiling analysis. *PLoS Comput Biol*, 11(11):e1004575, Nov 2015.

[12] Mads E Hauberg, John F Fullard, Lingxue Zhu, Ariella T Cohain, Claudia Giambartolomei, Ruth Misir, Sarah Reach, Jessica S Johnson, Minghui Wang, Manuel Mattheisen, Anders Dupont Børglum, Bin Zhang, Solveig K Sieberts, Mette A Peters, Enrico Domenici, Eric E Schadt, Bernie Devlin, Pamela Sklar, Kathryn Roeder, Panos Roussos, and CommonMind Consortium. Differential activity of transcribed enhancers in the prefrontal cortex of 537 cases with schizophrenia and controls. *Mol Psychiatry*, 24(11):1685–1695, 11 2019.

[13] Giovanni Iacono, Ramon Massoni-Badosa, and Holger Heyn. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol*, 20(1):110, 06 2019.

[14] Giovanni Iacono, Elisabetta Mereu, Amy Guillaumet-Adkins, Roser Corominas, Ivon Cuscó, Gustavo Rodríguez-Esteban, Marta Gut, Luis Alberto Pérez-Jurado, Ivo Gut, and Holger Heyn. bigscale: an analytical framework for big-scale single-cell data. *Genome research*, 28(6):878–890, 2018.

[15] Diego Adhemar Jaitin, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. Dissecting immune circuits by linking crispr-pooled screens with single-cell rna-seq. *Cell*, 167(7):1883–1896.e15, Dec 2016.

[16] Xin Jin, Sean K Simmons, Amy Guo, Ashwin S Shetty, Michelle Ko, Lan Nguyen, Vahbiz Jokhi, Elise Robinson, Paul Oyler, Nathan Curry, Giulio Deangeli, Simona Lodato, Joshua Z Levin, Aviv Regev, Feng Zhang, and Paola Arlotta. In vivo perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 370(6520), 11 2020.

[17] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell umi data. *Genome Biol*, 21(1):196, 08 2020.

[18] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[19] Lin Li, Hao Dai, Zhaoyuan Fang, and Luonan Chen. Ccsn: Single cell rna sequencing data analysis by conditional cell-specific network. *bioRxiv*, 2020.

[20] Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher, and Berthold Göttgens. Btr: training asynchronous boolean models using single-cell expression data. *BMC Bioinformatics*, 17(1):355, Sep 2016.

[21] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res*, 44(22):e164, 12 2016.

[22] David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

[23] Aivi T Nguyen, Kui Wang, Gang Hu, Xuran Wang, Zhen Miao, Joshua A Azevedo, EunRan Suh, Vivianna M Van Deerlin, David Choi, Kathryn Roeder, Mingyao Li, and Edward B Lee. Apoe and trem2 regulate amyloid-responsive microglia in alzheimer's disease. *Acta Neuropathol*, 140(4):477–493, Oct 2020.

[24] Tomasz J Nowakowski, Aparna Bhaduri, Alex A Pollen, Beatriz Alvarado, Mohammed A Mostajo-Radji, Elizabeth Di Lullo, Maximilian Haeussler, Carmen Sandoval-Espinosa, Siyuan John Liu, Dmitry Velmeshev, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*, 358(6368):1318–1323, 2017.

[25] Neelroop N Parikshak, Rui Luo, Alice Zhang, Hyejung Won, Jennifer K Lowe, Vijayendran Chandran, Steve Horvath, and Daniel H Geschwind. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, 155(5):1008–21, Nov 2013.

[26] Damon Polioudakis, Luis de la Torre-Ubieta, Justin Langerman, Andrew G Elkins, Xu Shi, Jason L Stein, Celine K Vuong, Susanne Nichterwitz, Melinda Gevorgian, Carli K Opland, et al. A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron*, 103(5):785–801, 2019.

[27] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*, 17(2):147–154, 02 2020.

[28] John E Reid and Lorenz Wernisch. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, 32(19):2973–2980, 2016.

[29] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[30] F Kyle Satterstrom, Jack A Kosmicki, Jiebiao Wang, Michael S Breen, Silvia De Rubeis, Joon-Yong An, Minshi Peng, Ryan Collins, Jakob Grove, Lambertus Klei, et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, 180(3):568–584, 2020.

[31] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):477, 2018.

[32] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 06 2019.

[33] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800, 2011.

[34] Dawn Thompson, Aviv Regev, and Sushmita Roy. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu Rev Cell Dev Biol*, 31:399–428, 2015.

[35] Jinjin Tian, Jiebiao Wang, and Kathryn Roeder. Esco: single cell expression simulation incorporating gene co-expression. *bioRxiv*, 2020.

[36] Dmitry Velmeshev, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H Rowitch, and Arnold R Kriegstein. Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, 364(6441):685–689, 2019.

[37] A Jeremy Willsey, Stephan J Sanders, Mingfeng Li, Shan Dong, Andrew T Tebbenkamp, Rebecca A Muhle, Steven K Reilly, Leon Lin, Sofia Fertuzinhos, Jeremy A Miller, Michael T Murtha, Candace Bichsel, Wei Niu, Justin Cotney, A Gulhan Ercan-Sencicek, Jake Gockley, Abha R Gupta, Wenqi Han, Xin He, Ellen J Hoffman, Lambertus Klei, Jing Lei, Wenzhong Liu, Li Liu, Cong Lu, Xuming Xu, Ying Zhu, Shrikant M Mane, Ed S Lein, Liping Wei, James P Noonan, Kathryn Roeder, Bernie Devlin, Nenad Sestan, and Matthew W State. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, 155(5):997–1007, Nov 2013.

[38] Shixiong Zhang, Xiangtao Li, Qiuzhen Lin, and Ka-Chun Wong. Review of single-cell rna-seq data clustering for cell type identification and characterization. *arXiv preprint arXiv:2001.01006*, 2020.

[39] Juan Zhao, Yiwei Zhou, Xiujun Zhang, and Luonan Chen. Part mutual information for quantifying direct associations in networks. *Proc Natl Acad Sci U S A*, 113(18):5130–5, May 2016.

[40] Lingxue Zhu, Jing Lei, Bernie Devlin, and Kathryn Roeder. Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *The annals of applied statistics*, 11(3):1810, 2017.

[41] Lingxue Zhu, Jing Lei, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019.